Supporting Information S2 for 'Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity'

Wilson J. Wright, Kathryn M. Irvine, Emily S. Almberg, and Andrea R. Litt

Descriptions of alternative false positive occupancy models

In our simulation study, the alternative false positive occupancy models we explored were based on summarizing the count detections to binary indicators (i.e., let $D_{ij\cdot k'} = \mathrm{I}(C_{ij\cdot k'} > 0)$) and/or only analysing data for the focal species. The resulting models are closely related to some other approaches that have been previously developed (Chambert, Miller, & Nichols, 2015; Chambert et al., 2018a; Chambert, Waddle, Miller, Walls, & Nichols, 2018b) and we provide more details on these comparisons here. The mathematical relationships between these alternative models is shown using a simple situation with two species ($K = 2$) and assuming the probabilities of occupancy and detection rates are constant across sites and visits. These connections illustrate how our approach provides a general framework for modelling false positive detections in occupancy models with each of these alternative models representing a special case. The connections between each model are based how the data are summarized and included in the analysis. Specifically, the approaches using binary data are cases where the observed data result from summarizing detection counts and the single species models assume the non-focal species is present at every site. The single species models may be appropriate for scenarios where false positive detections result from some omnibus source, as described by Chambert et al. (2015) and Chambert et al. (2018b). However, false positives from misidentified detections of another species are likely for many studies. For each model, we assume confirmed visits are available that allow the detection counts and corresponding species identifications to be directly observable for a portion of the visits. Each count and associated species $C_{ijkk'}$ is available at confirmed visits compared to the

unconfirmed visits where only the ambiguous counts, $C_{ij\cdot k'}$ are available. Note that in these comparisons we rely on the alternative model representation that is described in Supporting Information S1.

Two species, count detections

Here we outline the model described in main text section 2.2 for the specific case of $K = 2$ total species. Rewriting the model for this example helps make the comparisons to other models more clear. First, each species may be present or not at a site such that

$$Z_{i1} \sim \text{Bernoulli}(\psi_1),$$
$$Z_{i2} \sim \text{Bernoulli}(\psi_2).$$

For the unconfirmed visits, where only the ambiguous species classifications are available, the classified counts for each species are conditional on the occupancy states. In this case we have

$$[C_{ij\cdot 1} \mid Z_{i1}, Z_{i2}] \sim \text{Poisson}(z_{i1}\lambda_1\theta_{11} + z_{i2}\lambda_2\theta_{21}),$$
$$[C_{ij\cdot 2} \mid Z_{i1}, Z_{i2}] \sim \text{Poisson}(z_{i1}\lambda_1\theta_{12} + z_{i2}\lambda_2\theta_{22}).$$

This formulation means that the rate of ambiguous classifications for a species depends on which species are present at a particular site. If neither species is present, $Z_{i1} = Z_{i2} = 0$, then both rates are zero and no detections can be observed for either species. Modelling the data from confirmed visits allows the species detection rates and classification probabilities

to be estimated. For these data we model

$$[C_{ij11} \mid Z_{i1} = 1] \sim \text{Poisson}(\lambda_1 \theta_{11}),$$
$$[C_{ij12} \mid Z_{i1} = 1] \sim \text{Poisson}(\lambda_1 \theta_{12}),$$
$$[C_{ij21} \mid Z_{i2} = 1] \sim \text{Poisson}(\lambda_2 \theta_{21}),$$
$$[C_{ij22} \mid Z_{i2} = 1] \sim \text{Poisson}(\lambda_2 \theta_{22}).$$

Again the rates depend on the occupancy states, but the species associated with each detection is known for the confirmed visits.

Two species, binary detections

We explored a model that explicitly includes two species, but only incorporates binary detection data. This approach was described by Chambert et al. (2018a), however here we do not include the species interactions influencing the modelled probabilities. Instead, we assume that occurrence and detection of each species are both independent of the other species. For the binary detection data, let $D_{ijkk'}$ be defined by an indicator function for whether at least one detection from species k was classified as species k' for visit j to site i. That is, $D_{ijkk'} = \text{I}(C_{ijkk'} > 0)$, where I() denotes an indicator function returning 1 if the given argument is true and 0 otherwise. Similarly, let $D_{ij\cdot k'}$ be defined by an indicator function for whether at least one detection was classified as species k' for visit j to site i. For the unconfirmed visits, $D_{ij\cdot k'}$ represents ambiguous detections of species k' because a 1 could result from at least one true detection, at least one misclassified detection of the other species, or both.

As before, we model the presence of each species at a site as

$$Z_{i1} \sim \text{Bernoulli}(\psi_1),$$
$$Z_{i2} \sim \text{Bernoulli}(\psi_2).$$

For the unconfirmed visits, the probability of an ambiguous detection of species 1 depends on the occupancy state of both other species. We explicitly write out the distributions for ambiguous detections of species 1 below, and those for species 2 can be described similarly. There are four possible multi-species occupancy states and we model ambiguous detections of species 1 as

$$[D_{ij\cdot 1} \mid Z_{i1} = 0, Z_{i2} = 0] = 0 \text{ wp1},$$
$$[D_{ij\cdot 1} \mid Z_{i1} = 1, Z_{i2} = 0] \sim \text{Bernoulli}(p_{11}),$$
$$[D_{ij\cdot 1} \mid Z_{i1} = 0, Z_{i2} = 1] \sim \text{Bernoulli}(p_{21}),$$
$$[D_{ij\cdot 1} \mid Z_{i1} = 1, Z_{i2} = 1] \sim \text{Bernoulli}(p_{11} + p_{21} - p_{11}p_{21}).$$

Again, no detections are possible at sites where both species are absent. To connect the parameters in this model to those for count detections, we can use the probability of a count larger than zero from the Poisson distribution. Based on this, we define $p_{11} = 1 - \exp(-\lambda_1\theta_{11})$ representing the probability of at least one correctly classified detection for species 1. More generally, we have $p_{kk'} = 1 - \exp(-\lambda_k\theta_{kk'})$. Using these definitions, the relative activity rate can also be derived from the model. Note that $\lambda_k\theta_{kk'} = -\log(1 - p_{kk'})$ and $\lambda_k\theta_{k1} + \lambda_k\theta_{k2} = \lambda_k$ because $\theta_{k1} + \theta_{k2} = 1$. Then, λ_1 can be estimated as a function of the detection probabilities where $\lambda_1 = -(\log(1 - p_{11}) + \log(1 - p_{12}))$. When both species are present at a site, $Z_{i1} = Z_{i2} = 1$, the probability of an ambiguous detection corresponds to the complement of no true detections and no false detections. Mathematically, this is represented as $1 - (1 - p_{11})(1 - p_{21}) = p_{11} + p_{21} - p_{11}p_{21}$ for species 1.

At the confirmed visits, we again know the species associated with all detections and the corresponding classifications made for each. This allows the confirmed detections to be modelled as

$$
\begin{aligned}
[D_{ij11} \mid Z_{i1} = 1] &\sim \text{Bernoulli}(p_{11}),\\
[D_{ij12} \mid Z_{i1} = 1] &\sim \text{Bernoulli}(p_{12}),\\
[D_{ij21} \mid Z_{i2} = 1] &\sim \text{Bernoulli}(p_{21}),\\
[D_{ij22} \mid Z_{i2} = 1] &\sim \text{Bernoulli}(p_{22}).
\end{aligned}
$$

Each of these distributions is related to those for the counts at confirmed visits described by the model in section S2.1. All are equivalent to modelling the probability of at least one count based on the Poisson distribution.

Single-species, count detections

Next, consider the case where only data for the focal species (species 1) is included and the corresponding detection counts are modelled. In this case, all data corresponds to detections that are classified as species 1 but these can also include false positives resulting from misclassified detections of species 2. This approach closely aligns with the model described by Chambert et al. (2018b) but with two main differences. First, they specifically modelled the probability of zero detections separately from the counts using a hurdle model. Alternatively, we keep the probability of species 1 going undetected as described by the Poisson distribution corresponding to a count of zero detections. Second, for the confirmed visits, they described a scenario where individual detections were randomly sampled and represented this process with a Hypergeometric distribution in their model. We describe the scenario where all detections from confirmed visits are verified and therefore do not include this component of the model. These could be important aspects to account for in particular

studies, but we ignore them here to more clearly show the connections with our general model framework.

If false positives for the focal species can only occur due to misclassified detections of species 2, this model assumes that species 2 is present at every site ($\psi_2 = 1$). Then, occupancy for species 1 is modelled as

$$Z_{i1} \sim \text{Bernoulli}(\psi_1),$$

while $Z_{i2} = 1$ with probability 1 (hereafter 'wp1'). For the unconfirmed visits (where species identifications are ambiguous), the rate of observed counts classified as species 1 depends on whether the site is occupied by species 1 or not. Note, however, that there can always be a contribution of false positive detections from species 2 because it is present at every site. Therefore, the counts from unconfirmed visits are modelled as

$$\begin{aligned}[C_{ij\cdot 1} \mid Z_{i1} = 1] &\sim \text{Poisson}(\lambda_1\theta_{11} + \lambda_2\theta_{21}),\\ [C_{ij\cdot 1} \mid Z_{i1} = 0] &\sim \text{Poisson}(\lambda_2\theta_{21}).\end{aligned}$$

We separated these conditional distributions here, but this is equivalent to the distribution for $C_{ij\cdot 1}$ in section S2.1 under the assumption that $Z_{i2} = 1$ wp1. Similarly, the model for counts at confirmed visits is also identical to the model in section S2.1 under this assumption. We retain the notation indicating that false positive detections specifically result from species 2, but more generally this could be described as an omnibus source of false positives that is present at every site. We then model the detections classified as species 1 at confirmed visits with

$$\begin{aligned}[C_{ij11} \mid Z_{i1} = 1] &\sim \text{Poisson}(\lambda_1\theta_{11}),\\ C_{ij21} &\sim \text{Poisson}(\lambda_2\theta_{21}).\end{aligned}$$

Our formulation shows that the rate of correctly classified detections from species 1 (C_{ij11}) is a combination of the overall detection rate (λ_1) and the probability of correctly classifying the species (θ_{11}). Unlike the previous model, however, these parameters are not estimated uniquely and we only have information about their product. This can be seen in the model described by Chambert et al. (2018b) which has a rate parameter for true detections (represented by $\lambda_1\theta_{11}$ here) at occupied sites and a rate parameter for the false detections (equal to $\lambda_2\theta_{21}$) which occur at all sites. If the model described here also included the incorrectly classified detections of species 1 (C_{ij12}), the detection rate and classification probabilities could both be estimated.

Single-species, binary detections

The final model we make comparisons to is the 'observation confirmation' model described by Chambert et al. (2015). As with the previous single species model (section S2.3), if false positive detections for species 1 result from the erroneous classification of detections for species 2, this approach must assume that species 2 is present at every site. We only include detections classified as species 1 in the analysis. In that case, the model for occupancy is again

$$Z_{i1} \sim \text{Bernoulli}(\psi_1),$$

while $Z_{i2} = 1$ wp1. At the unconfirmed visits, we model the ambiguous detections of species 1 as

$$[D_{ij\cdot 1} \mid Z_{i1} = 1] \sim \text{Bernoulli}(p_{11} + p_{21} - p_{11}p_{21}),$$
$$[D_{ij\cdot 1} \mid Z_{i1} = 0] \sim \text{Bernoulli}(p_{21}).$$

Here false positive detections are possible at every site because species 2 is always present.

This distribution is analogous to that for the confirmed visits described by the single-species, count detections model (section S2.3) based on the probabilities of at least one detection classified as species 1. Finally, the model for the confirmed visits is represented as

$$[D_{ij11} \mid Z_{i1} = 1] \sim \text{Bernoulli}(p_{11}),$$
$$D_{ij21} \sim \text{Bernoulli}(p_{21}),$$

which is also analogous to the model described in section S2.3. For this model, however, we estimate $\lambda_1 = -\log(1 - p_{11})$ because p_{12} is not included in this model. Consequently, this approach only has information about $\lambda_1\theta_{11}$ and is unable to directly estimate the relative activity rate.

REFERENCES

Chambert, T., Campbell Grant, E. H., Miller, D. A. W., Nichols, J. D., Mulder, K. P., & Brand, A. B. (2018a). Two-species occupancy modeling accounting for species misidentification and nondetection. *Methods in Ecology and Evolution*, *9*(6), 1468–1477.

Chambert, T., Miller, D. A. W., & Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, *96*(2), 332–339.

Chambert, T., Waddle, J. H., Miller, D. A. W., Walls, S. C., & Nichols, J. D. (2018b). A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing. *Methods in Ecology and Evolution*, *9*(3), 560–570.